WP1 – Modelling Phagocytosis

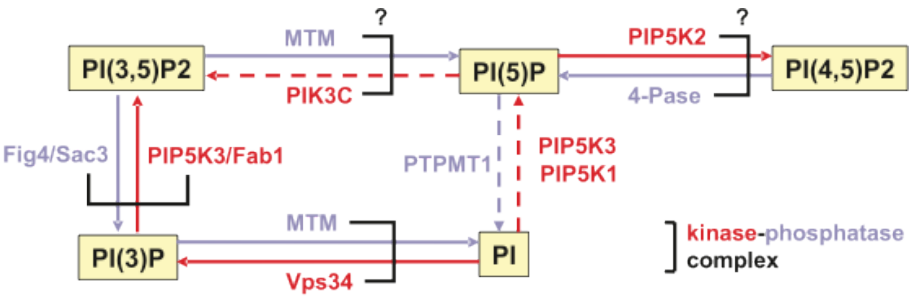
The overall goal of WP1 was to develop models of phagocytosis, based on existing expert knowledge (ICTSM) and on data (JSI); both existing data and data collected within the project were to be used for this purpose. In addition, methods for automated modelling that can make use of both data and expert knowledge were to be further developed to meet the typical needs of systems biology modelling tasks (JSI).

Major progress was made towards achieving these goals, but some changes were inevitable due to unforeseeable circumstances. At ICTSM, WP1 experienced problems that affected the progress of the project including long-term illness and subsequent tragic death of the mathematician Professor Jaroslav Stark, and the late start and early departure of the research fellow at ICSTM, Dr Barbara Szomolay (April 2009-October 2011). A biomathematician, Dr. Vahid Shaherezei, was recruited to the project and managed the ICSTM side of WP1. Due to this, the scope of the models constructed is more limited than originally foreseen.

At JSI, the scope of WP1 was extended to analysing the majority of data generated within the project. It was originally planned to use time-course data on protein concentrations to learn models of the dynamics of phagocytosis, however the limited availability of this kind of data meant we were not able to do as much of this as originally foreseen. Since the majority of the data generated in the project included different types of screens (genome or compound screens, flow-cytometry or image-based), a large amount of effort was devoted to integrating and analysing these, which is reflected in the outcomes of the project.

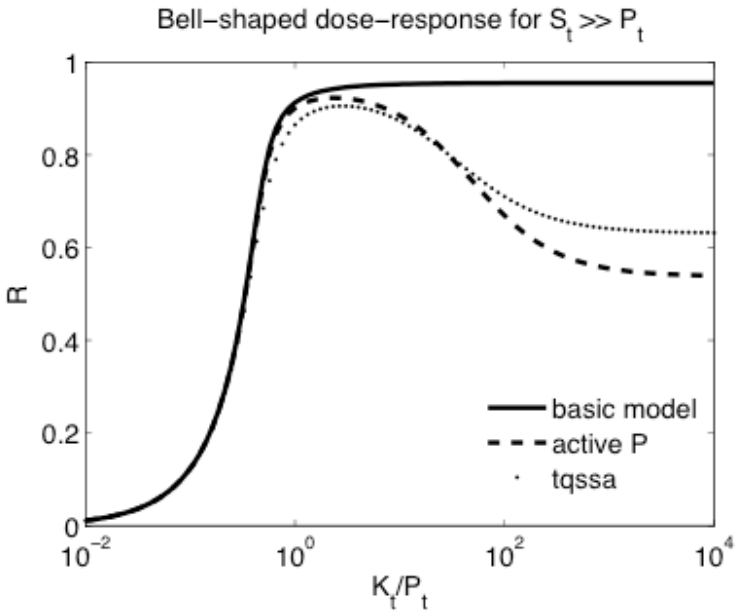
At ICSTM, we first compiled a signalling network for phagosome maturation based on an extensive literature survey. Based on the complexity of the initial network and the limited available data, we concluded that construction of a comprehensive mathematical model of this system is not yet feasible. We then decided to focus on particular aspects of this system and model them in detail.

Phagosome maturation is regulated by the key tagging molecules Rab proteins and phosphoinositides (PIPs). Inspecting the signalling network that regulates PIPs revealed a surprising interaction between antagonistic enzymes, including interactions between Vps34 kinase-myotubularin phosphatases (see Figure WP1-1).



Legend Figure WP1-1: Identified or hypothesized kinase-phosphatase complexes (PIK3C – class I PI 3-kinase, PTMT1 – PTEN-like phosphatase). Full lines represent in vivo, dashed line in vitro data.

Using mathematical modelling we investigated the role of these interactions, and found that the complex formation can produce novel forms of switch-like and bell-shaped response (Figure WP1-2). We have postulated that these could have a functional role in the temporal regulation of PIPs during phagosome maturation (Szomolay and Shahrezaei, 2012).



Legend Figure WP1-2: Bell-shaped dose response for the active phosphatase case in the zero-order regime. Kt, Pt, St are the total kinase, phosphatase, and substrate concentrations, respectively. R denotes the fraction of total phosphorylated substrate. The dashed line represents the response curve for the extended model in the active phosphatase case and the solid line represents the response curve for the corresponding basic model of a phosphorylation-dephosphorylation cycle.

We have been also involved in statistical analysis of multiparametric imaging data (Collinet et al., 2010) from our partners at MPI-CBG, describing the

intensity of EEA1 and APPL1 endosomes Rab5 effector proteins necessary for endosomal trafficking) at the single endosome level for wild type and a range of RNAi knockdowns. In addition to MATLAB, image analysis software (Motion Tracking, developed by Y. Kaladzidis in the Zerial lab, Dresden) was used to analyse the data. A recent study has shown that APPL1-positive and EEA1-positive endosomes have a transient overlap (APPL1 endosomes represent an earlier stage than EEA1 endosomes) and that the switch from the APPL1 to the EEA1 stage is controlled by an important phospholipid, PI(3)P (Zoncu et al., 2009). One goal is to show that the single endosome data is consistent with this finding, but so far we have not found strong evidence for this. In addition we are looking at the heterogeneity of the markers and how this is regulated. We hope to add insights into the mechanisms of robustness in endosome trafficking. This is still an ongoing work in collaboration with MPI-CBG and we plan to publish these results as soon as all the analyses are finished.

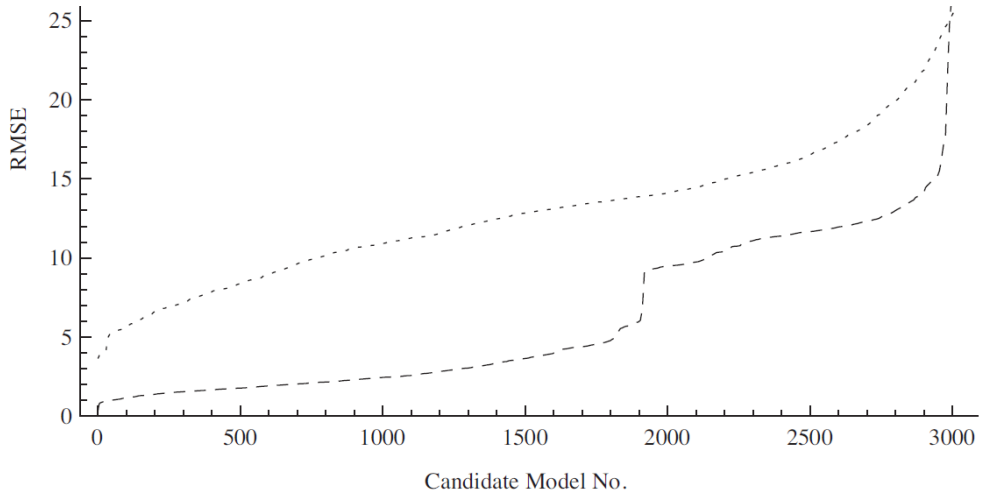
At JSI, we pursued two major lines of research. First, we continued the development of methods for the automated modelling of dynamic systems (in the form of ordinary differential equations; ODEs) that can make use of both experimental data and expert knowledge, and so meet the needs of typical systems biology modelling tasks. We also investigated the use of these automated methods for modelling endosome maturation and LDL trafficking, where time-course data on protein concentrations was available. Second, we applied machine learning methods for structured output prediction, and in particular predictive clustering, to a large number of datasets generated from high-throughput screens related to phagocytosis of intracellular pathogens carried out by the project partners.

We have made significant advances in machine learning methods for automated modelling of dynamic systems (Cerepnalkoski et al., 2012), not only relevant to the automated modelling of phagocytosis but also addressing other modelling tasks in systems biology. First, an extended formalism for encoding domain-specific knowledge for automated modelling was developed. It is based on entities and processes, and modelling templates thereof: it facilitates the grouping of entities and processes into compartments, an important concept in systems biology. Second, a new software platform was implemented that supports the learning of models in this formalism. This software platform allows for the use of different local and global optimization methods for parameter estimation in ODE models. It also allows t h e u s e of different objective functions (in addition to the sum of squared errors) within these optimization methods.

The utility of the improved methods was evaluated within the simple context of estimating the parameters of a single model structure and a more challenging setting of selecting an appropriate model structure. For the single model structure, we used the model of endosome maturation proposed by Del Conte-Zerial et al (2008). We compared the performance of local derivative-based and global meta-heuristic optimization methods for parameter estimation under different observation scenarios and varying level of noise in the data. The scenarios cover a wide range of situations, from the simplest

one of complete observability, where the concentrations of all state variables are assumed to be directly measurable, to the most complex (and realistic) scenario, where the observations are linear combinations of state variables (total concentration of protein active and passive states). Global meta-heuristic methods clearly and significantly outperform the local derivative-based method. These results hold for both real and artificial data, for all observation scenarios considered, and for all amounts of noise in the data (Tashkova et al., 2011).

We also considered the use of different optimization methods in the context of estimating the parameters in a large number of model structures, which is typical for automated modelling approaches (Cerepnalkoski et al., 2012). We compare a global and a local optimization method, as used in the context of the automated modelling tool, on four modelling tasks, each involving a large number of model structures. We considered the difference in performance of the best models found, the overall quality of all models considered, and each individual model considered; across all of these, the global optimization method used in the context of automated modelling performs much better than the local optimization method. For illustration, consider the error profiles of the models with parameter values fitted by each of the two methods, shown in increasing error order (Figure WP1-3): These clearly show the superiority of global optimization. The use of global optimization for parameter estimation eliminates a major weakness of automated modelling methods that we identified during the first phase of the project.



Legend Figure WP1-3: Error profile curves for ODE model structures whose parameters have been fitted by a local (top curve) and a global (bottom curve) optimization method. Much better models are found when using global optimization methods for parameter fitting.

With the new software platform ready, in collaboration with MPI-CBG Dresden we formulated a library of entity and process templates for modelling Low-Density-Lipoprotein (LDL) trafficking, based on domain knowledge. Time-course profile data of LDL concentrations in different compartments (vesicular/non-vesicular) and stages (before/on/after) Rab5 were also provided, measured under stimulation (10 min. pulses) with different

concentrations of EGF (from 30 to 3000 ng/ml). Good fit of the model with the estimated parameters to the measured data was achieved, both for the case of a single model structure and several alternative model structures.

In the first half of the project we started using different machine learning methods to analyse data provided by the PHAGOSYS partners, with the aim of gaining a better understanding of the process of phagocytosis. In particular, we used methods for structured output prediction, such as predictive clustering trees, to analyse a variety of data related to phagocytosis of mycobacteria. These included microarray data with time course profiles of gene expression levels in response to infection: Type 1 and Type 2 human macrophages were infected with M. tuberculosis, and Schwann cells were infected with M. leprae (data from LUMC). Interesting groups of genes with common functional annotations and clearly defined response patterns were identified for Schwann cells. This enabled the formulation of a hypothesis about a new candidate gene, previously unknown in the regulation of intercellular growth of mycobacteria. The hypothesis was subsequently investigated in wet-lab experiments and silencing of the gene was found to reduce bacterial load significantly, as predicted.

We have then used predictive clustering trees, as well as feature ranking methods, to analyse data from several screens collected by different PHAGOSYS partners. These included an siRNA image-based screen of endocytosis in HeLa cells (data from MPI-CBG), a flow cytometry screen and a smaller image-based siRNA screen aimed at studying MHC Class II antigen presentation (data from NKI), which we tried to inter-relate by our analyses. Finally, we analysed data from a large compound screen (from LUMC) of MelJuSo cells infected with M. tuberculosis and HeLa cell infected with S. typhimurium.

We analysed data from a genome-wide siRNA image screen performed at MPI-CBG to study the process of endocytosis in the HeLa cell line. The phenotype images obtained after silencing each gene were analysed using MotionTracking (MPI-CBG in-house software) to produce feature-based descriptions of the images. We described each gene with its functional annotation (using GO terms). The goal of the study was to provide knowledge about endocytosis and to annotate the hypothetical genes with possible functions using the methodology developed by Kocev (2011).

The NKI provided data from two screens concerning MHC Class II antigen presentation. The first screen includes flow cytometry data for 16675 genes, silenced in the MelJuSo cell line, where two antibodies, L243 and CerCLIP, were observed. Then, 276 candidate genes from this screen (selected by a z-score cut-off) were used for the second screen, where images of the cells were taken and phenotypes analysed with the CellProfiler software to extract features. Using (semi)manual clustering on these features, 19 bins were identified (Paul et al., 2011), but a number of genes/images were left unassigned. We described each gene with its GO annotations and 13 CellProfiler features.

We performed several analyses in order to merge the data from these three studies (MPI-CBG image screen and the NKI's flow cytometry and image screens) and identify some false negatives, possibly resulting from the z-score cutoff in the flow cytometry study. More precisely, we aimed to find genes (from the MPI-CBG genome wide study) that were not selected for in the NKI imaging study, due to the stringent definition of a hit (a z-score larger than 2 or smaller than -2). We conducted two types of analyses: gene set enrichment using SEGS-BioMine (Podpecan et al., 2011), and then predicted the possible assignment of the genes from MPI-CBG data to the bins from the NKI study. Using gene set enrichment, we discovered gene functions for the highly ranked genes from the flow cytometry study, dependent on the antibody used.

The second analysis investigated the possibility of merging the three studies, hoping that the analysis performed in the genome-wide study at MPI-CBG could help elucidate some genes that were found to be negatives in the NKI flow cytometry study. The results showed that the information about the phenotypes from the MPI-CBG study cannot be (directly and in a straight-forward manner) connected to the phenotype bins defined by the NKI phenotype study. There are several factors that contribute to this outcome: the use of different cell-lines (HeLa versus MelJuSo), the different cellular mechanisms studies (endocytosis versus antigen processing), and different imaging and image description techniques (MotionTracking versus CellProfiler). The integration of the screen data would be more successful if a smaller subset of genes from the NKI study was selected, and the closest matches from the MPI-CBG study calculated. This is the topic of an ongoing collaboration between JSI and NKI.

For the third analysis, we performed feature ranking on t he t hree data representations. We used RReliefF as a feature ranking algorithm (Robnik-Šikonja and Kononenko). Namely, we would like to obtain an ordered list of genes or gene functions that are most relevant for the decrease/increase of bacterial load. After that, using the top 40 ranked genes we constructed a gene network using the STRING database.